Response to Referee #3

Referee comments are in plain text
Author comments are in BOLD
Manuscript text is in *italics*
Added Text is in ***Bold Italics***
Line numbers refer to the marked manuscript

We thank the Referee for the comments and address each point below.

This is a well written paper that proposes a new methodology to predict total and infragravity swash elevation. As such it is of interest to NHESS and coastal scientists/practitioners. The methodology followed is correct and well explained. In particular, there is a very clear explanation of Genetic Programming and how this technique has been used for this work. This is very well written in a way which is suitable for non-experts approaching the methodology for the first time. The data used are of very good quality and there is a good explanation of the range of parameters covered by the dataset. The results are discussed in concise and detailed way and the accuracy improvement over existing relationships is demonstrated.

My only minor suggestion is that the use of both MSE and RMSE is redundant and one of the two can be omitted. Therefore I recommend publication after minor revision.

We prefer to keep both MSE and RMSE in the manuscript to aid in the rapid comparisons with future prediction schemes — for instance, a future study may report only MSE or RMSE.

Minor corrections/suggestions.
Abstract Line 14: change the sentence: it contributes to the error, maybe it is contributes to the reduction of the error. However, beware of repetitions.

We addressed this.

Line 14: ***"****Using this newly compiled dataset we demonstrate that a ML approach can reduce the prediction errors compared to well-established parameterizations and therefore* ***it may improve*** *coastal hazards assessment (e.g. coastal inundation)."*

Also many repetitions of "wave runup" in the introduction, try to rephrase.

Line 33

We changed "*The first predictors of wave runup were…***" into "***The first predictors of* ***these phenomena*** *were…"*

Line 123, concomitant is possibly better replaceable by "associated".

Line 126: We replaced "*concomitant***" with "*** ***associated"***

Line 143-147: specify the countries of the beaches named as not all authors might be familiar with these.

We added the information about the countries where the experiments were performed.

Line146-150: "*The dissipative beaches of the original dataset (Fig. 2 d, h) are Terschelling* ***(Netherlands)*** *and Agate* ***(USA)****, and for the new dataset Ngarunui* ***in New Zealand*** *(although, during the experiment, the beach also experienced intermediate conditions). The purely intermediate beaches for the original and new dataset are Scripps* ***(USA)*** *and TrucVert (****France)****. Some beaches of the original dataset* ***(USA)*** *represent both intermediate and reflective conditions: Duck 94, Gleneden, Sandy Duck, Delilah and Duck 82. San Onofre for the original and Tairua* ***(New Zealand)*** *for new dataset are reflective.*"

Line 170. You might want to specify which is your stopping criterion, and when do you consider the solutions stable.

We moved and clarified the sentence from lines 203 – 204 to line 173-175

Line 173-175: "***The search is stopped after the GP evaluated 10^{11} formulas because the solutions stabilized and no significant improvement in formula performance was found."***

Line 237: overfitting is mentioned, but it could be useful to explain what this is in the present context. Explanation in 240 occurs after the first use of the term and it is not clear.

We define overfitting before mentioning it (moved to line 242) and we add a definition of overfitting from a new reference: Dietterich T.: Overfitting and Undercomputing in Machine Learning, ACM Comput. Surv., 27 (3), doi:10.1145/ 212094.212114 1995.

Line 242-248: *"Generally, extremely complicated predictors fit the training and validation dataset better than simpler predictors but they may lose generalization power* ***when tested on a separate testing dataset*** *(overfitting).* ***In other words a predictor with overfitting could represent the noise in the training and validation subsets instead of defining a general predictive rule (e.g., Dietterich, 1995) and therefore it will result in smaller training errors but in higher testing errors.*** *Several viable techniques exist for selecting the best solution to avoid overfitting, all meant to balance the fact that simpler solutions (the minimum description length) might risk losing more accurate information contained in more complex models (e.g., O`Neill et al., 2010)."*

Some sentences are written in present tense (e.g. we use at the beginning of Section 3.3, and "...finally selected" at the start of 4.2). Please make the tense consistent.

We changed the past tense into present tense.

Line 127, 128

We changed "*were calculated****" to "******are calculated******" and "***were located****" to "******are located******"**

Line 192

We changed "*We searched"* **to "*We searched*"**

Line 223, 225

We changed "*we used"* **to "*we use*", We changed "***was tested"* **to "*is tested*",**

Line 256

We changed "*selected"* **to "*select*"**

Line 374-375

We changed "*did not"* **to "*do not*", and "***found*" **to "*find*"**

Also, in Line 314 it is mentioned that experiments in Ngarunui beach are carried out under mild dissipative conditions. Is the difficulty in predicting these results due to the particular combination of H and T (hence L)? It would be useful to be more detailed in explaining this.

This sentence is connected to the previous one where we discussed that the Stockdon et al. formula for S_{Ig} has more scatter for Terschelling and Agate (which are dissipative). We did not highlighted the characteristic of these beaches in the paragraph, but we already defined them dissipative at line 277 (because the same happen for S_{tot}). Our intention on line 323 was to highlight the performance of the predictors on the dissipative beaches — settings where infragravity motion has the greatest importance. Ewe now clarify this on line 314-329.

Line 314-329: *"Generally the three formulas seem to perform similarly. Some differences are found in* ***dissipative settings (i.e.,*** *Agate and Terschelling)* ***—predictions by*** *Stockdon et al., (2006)* ***tend to overestimate S_{ig} compared to*** *the GP predictors* **.** *The same difficulty in predicting swash excursion on a dissipative beach is observed on Ngarunui (Fig. 7).* ***Even though*** *this experiment was performed under mild wave conditions (H_0 ~ 0.6-1.26 (m) and T_p ~8.1-12.4 (s), Table 1) compared to the experiments at Agate and Terschelling.* ***Note that dissipative beaches are the one were the infrragravity motion has greater importance****. Also Truc Vert presents dissipative conditions in the swash zone, while the surf zone is intermediate (ξ_0 up to 0.87 as reported by Senechal et al., 2011). For this experiment Eq. (13) and (7) (Fig. 7 a, c) overestimate S_Ig while Eq. (14)* ***has better performance for the dissipative beach Ngarunui,*** *suggesting that it could be the most appropriate for S_Ig predictions."*

In Line 356 it is claimed that the procedure followed is different from the use of a single data set. This needs clarification, as you always build one dataset that is divided in three for training validation and testing. The same was done in the development of ANN tools for overtopping in the CLASH project (van Gent et al. 2007), for example, when the dataset used was actually a composite one resulting from many datasets.

References

van Gent, M.R., van den Boogaard, H.F., Pozueta, B. and Medina, J.R., 2007. Neural network modelling of wave overtopping at coastal structures. Coastal Engineering,

54(8), pp.586-593.

We now clarify our study — discussing how our method is different than previous data splitting work.

Line 368-372: "*In this work we use data compiled by Stockdon et al., (2006) to build new predictors, by the use of GP, for both total and infragravity swash elevations. We then test the generalizability of these new predictors using new data (including some extreme conditions). This is different from* ***many previous applications of ML in coastal settings in two ways: First, we are testing the ML-derived predictor on data that is collected from a different setting (compared to the training data)— three beaches not included in the training data. Second, the testing data includes events that are outside the data range of the training data — we are extrapolating the ML-derived predictor as a test of its generalizability.***

Thank you for considering the revised version of this manuscript for publication in NHESSD